## PATENT ABSTRACTS OF JAPAN

(11)Publication number:

11-066196

(43) Date of publication of application: 09.03.1999

(51)Int.CI. G06F 19/00
G06F 17/21
G06T 7/00
G06K 9/20
H04N 1/21
H04N 1/40

(21)Application number: 09-220426

15.08.1997

(71)Applicant: RICOH CO LTD

(72)Inventor: SAITO TAKASHI

**ABE TEI** 

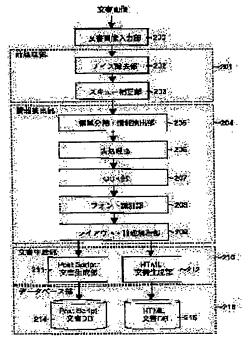
**KOUCHI TSUKASA** 

(54) DOCUMENT IMAGE RECOGNITION DEVICE AND COMPUTER-READABLE RECORDING MEDIUM WHERE PROGRAM ALLOWING COMPUTER TO FUNCTION AS SAME DEVICE IS RECORDED

## (57) Abstract:

(22)Date of filing:

PROBLEM TO BE SOLVED: To generate documents in various format complying with usages such as a document whose reproduction is given priority and a document whose contents are made important. SOLUTION: This device is equipped with a document image input part 200 which inputs a document image generated by optically reading a paper document, a preprocess part 201 which performs noise removal and a skew correcting process for the inputted document image, an information extraction part 204 which performs a recognizing and extracting process for a character area including character strings and/or a an image area including images of a graph, a table, a photograph or the like, and a character recognizing process for the character strings in the extracted character area and also analyzes the layout of the document image to extract layout information, a document generation part 210 which generates a PostScript document and an HTML document



according to the character recognition result and layout information extraction result, and a data base part 213 which stores the generated PostScript document and HTML document respectively.

## LEGAL STATUS

[Date of request for examination]

13.06.2002

[Date of sending the examiner's decision of

04.02.2003

rejection]

[Kind of final disposal of application other than the examiner's decision of rejection or application converted registration]

[Date of final disposal for application]

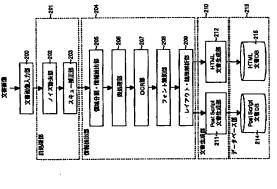
57.22 G 9/20 340 C 1/21 538 A 15/70 330 Q	(全16月)最終頁に続く000006747株式会社リコー	東京都大田区中馬込1丁目3番6号 商職 高志 東京都大田区中馬込1丁目3番6号 株式会 社リコー内 阿部 馀	東京都大田区中馬込1丁目3番6号 株式会社リコー内 社リコー内 幸地 司 東京都大田区中馬込1丁目3番6号 株式会 東リコー内 弁理士 酒井 宏明
F1 G06F1 G06K H04N G06F1	(71) 出願人	(72) 発明者(72) 発明者(72) 発明者	(72) 発明者 (74) 代理人
韓別記号 19/00 17/21 7/00 9/20 3 4 0 1/21	密金請求 未請求、請求項の数4 特額平9-220426	平成9年 (1997) 8月 15日	
(51) Int. Cl. 6 G 0 6 F G 0 6 T G 0 6 K H 0 4 N	(12) 田蘭器	(22) 出版日	

[64] 【死明の名称】文書画像認識装置およびその装置としてコンピュータを機能させるためのプログラムを記録したコン ピュータ競み取り可能な記録媒体

(57) [要約]

を重視した文書等、利用目的に応じた様々な形態の文書 【映題】 紙文書の再現を優先した文書や紙文書の内容 を生成すること。 紙文音を光学的に読み取ることによって 生成した文書画像を入力する文書画像入力部200と、 [解决手段]

t S c r i p t 文書およびH T M L 文書をそれぞれ格构 入力した文哲画像に対してノイズ除去およびスキュー柏 正処理を行う前処理部201と、文書画像から文字列を む画像領域の認識・抽出処理、抽出した文字領域の文字 列についての文字認識処理、および文書画像のレイアウ トを解析し、レイアウト情報の抽出処理を行う情報抽出 果に基づいて、PostScript文書およびHTM 含む文字領域および/または図、表,写真等の画像を含 第204と、文字認識結果およびレイアウト情報抽出結 L文書を生成する文書生成部210と、生成したPos するデータベース第213と、を値えている。



ひ文字領域および/または図、投、写真等の画像を含む 前記領域認識手段で抽出した文字領域の文字列について 【謝求項1】 紙文書を光学的に読み取ることによって **前記入力手段を介して入力した文豊画像から文字列を含** 画像領域を認識して抽出する領域抽出手段と、 主成した文督画像を入力する入力手段と、 欠字認識処理を行う文字認識手段と、 特許請求の範囲]

**竹記領域抽出手段の抽出結果に基づいて、前記文書画像** のレイアウトを解析し、レイアウト情報を抽出するレイ アウト情報抽出手段と、

前記文字認識手段による文字認識結果およびレイアウト て、ページ記述書籍を用いた第1の文魯を生成する第1 情報抽出手段によるレイアウト情報抽出結果に基づい の文書生成手段と、

前記文字認識手段による文字認識結果およびレイアウト て、構造化記述言語を用いた第2の文盤を生成する第2 **備報抽出手段によるレイアウト情報抽出結果に基づい** の文書生成手段と、

前記第1および第2の文書生成手段で生成した第1およ び第2の文書をそれぞれ格納する格納手段と、 を備えたことを特徴とする文替画像認識装置。

2

[讃求項2] 前記第1の文書は、PostScrip t形式またはPDF形式によって表現された文書である 【請求項3】 前記第2の文書は、SGML, HTML またはXMLによって表現された文書であることを特徴 ことを特徴とする請求項1に記載の文書画像認識装置。 とする請求項1に記載の文譽画像認識装置。

30 の文書画像認識装置の各手段としてコンピュータを機能 【請求項4】 前記請求項1~3のいずれか1つに記載 させるためのプログラムを記録したことを特徴とするコ ンピュータ読み取り可能な記録媒体。 [発明の詳細な説明]

[000]

した文書等、利用目的に応じた様々な形態のコンピュー 説み取ることによって得た文替画像から文字コードを抽 り、紙文書の再現を優先した文器や紙文書の内容を重視 た文魯画像認識装置およびその装置としてコンピュータ を機能させるためのプログラムを記録したコンピュータ [発明の属する技術分野] 本発明は、紙文啓を光学的に **タ上で利用可能な文書を生成することができるようにし** 出するという単なる文字認識処理を行うだけではなく、 紙文箏の持つ様々な情報を抽出して利用することによ 競み取り可能な記録に関する。

[0001]

[0002]

2 低の形で保存・流通されている文書は未だ多量に存在し することとなった。ところが、抵は閲覧性が良い等の利 の急速な普及は、オフィスにおける文書の電子化を促進 [従来の技術] パーソナルコンピュータやネットワーク 便性があるため、紙文書の生産は未だ止むことがなく、

アウト情報抽出手段によるレイアウト情報抽出結果に基 プいて、ページ記述言語を用いた第1の文書を生成する

特開平11-66196

ij

ていることが現状であり、このギャップがオフィスワー クにおける生産効率の低下の一因となっている。

上のデータという形で存在しているのは、相互の変換コ れた形とコンピュータ上のデータという形で存在してお 再利用といった一連の流れにおいて、文書は紙に記録さ り、このように文書が紙に記録された形とコンピュータ [0003] つまり、文敬の作成、流通、閲覧、管理、 ストが高いということが原因となっている。

により、紅文書をスキャナ等で読み取って文書画像を生 [0004]上記問題を解決する手段の一つとして、O 成した後、文督画像中の文字列を文字コードに変換する er)を挙げることができる。このOCRを用いること CR (Optical Character Read ことができる。 2

[0005]

ることは可能であるが、元の紙文書のレイアウト等まで た文書は単なるテキストの騒列であって、元の文语画像 う問題点があった。後育すれば、文掛は、作成者の意図 OCRを用いただけでは、文哲のレイアウトを抽出する ことができず、元の文むのレイアウトを再現した文也を 所たに生成したり、元の文書のレイアウトを利用して新 抽出することができないため、OCRによって生成され が持つ様々な情報を十分に利用することができないとい に応じて様々なレイアウト処理が施されているが、単に たなレイアウトの文書を生成したりすることはできなか CRは、文書画像中の文字列を文字コード情報に変換す [発明が解決しようとする標題] しかしながら、上紀O

を行うだけではなく、紙文母の持つ様々な情報を抽出し [0006] 本発明は上記に鑑みてなされたものであっ て、紙文費を光学的に脱み取ることによって得た文書画 像から文字コードを抽出するという単なる文字認識処理 て利用することにより、紙文봡の再現を優先した文むや な形態のコンピュータ上で利用可能な文書を生成するこ 低文書の内容を重視した文書等、利用目的に応じた様々 とができるようにすることを目的とする。

文字認識処理を行う文字認識手段と、前記領域抽出手段 **折し、レイアウト情報を抽出するレイアウト情報抽出手** 読み取ることによって生成した文母画像を入力する入力 手段と、前記入力手段を介して入力した文啓画像から文 字列を含む文字領域および/または図,表,写真等の画 前記領域認識手段で抽出した文字領域の文字列について の抽出結果に基づいて、前記文苷画像のレイアウトを解 段と、前記文字認識手段による文字認識結果およびレイ 像を含む画像領域を認識して抽出する領域抽出手段と、 め、請求項1の文書画像認識装置は、紙文費を光学的 [概題を解決するための手段] 上記目的を違成するた

3

特開平11-66196

第1の文格生成手段と、前記文字認識手段による文字認 館結果およびレイアウト情報抽出手段によるレイアウト 情報抽出結果に基づいて、構造化記述書語を用いた第2 の文哲を生成する第2の文都生成手段と、前記第1およ び第2の文哲生成手段で生成した第1および第2の文章 をそれぞれ格納する格納手段と、を個えたものである。 【0008】また、精束項2の文豊画像認識装置は、請 次項1に記載の文語画像認識表置に、請 次項1に記載の文語画像認識表置に、前記第1の 文语が、PostScIpt形式

よって表現された文化であるものである。 【0009】また、加水切3の文格回像認識装置は、請水項1に記載の文格画像認識装置において、前記第2の文格が、SGML、HTMLまたはXMLによって表現された文化であるものである。

[0010] さらに、請求項4のコンピュータ競み取り可能な記録媒体は、前記請求項1~3のいずれか1つに記載の文費函像認識装置の名手段としてコンピュータを機能させるためのプログラムを記録したものである。

「現内の状態の形態」以下、本発明の文華画像認識装置 およびその設留としてコンピュータを機能させるための プログラムを記録したコンピュータ総み取り可能な記録 媒体の実態の形態について、落付の図面を参照しつつ詳細に説明する。

(0012) (実施の形態1) 実施の形態1の文母画像 20維数配は、紅文母を光学的に結み取って生成した文母画像を入力し、入力した文母画像に基づいて、コンピュータ上で利用可能な文母を生成するものである。後言すれば、オフィスワーク等においては、文母を生産し、液 ゴレ、閲覧し、管理し、利用するという一型の流れがあ 3 ため、実施の形態1の文母画像認識装置は、上記流過および閲覧を考慮した文母と、利用を考慮した文母を生成することができるようにしたものである。

**再現することができるようにするために、実施の形態1** ptに代えて、PDF (Portable Docum は、文街作成者の意図をできるだけ正確に伝えることが 報まで忠実に再現することを目的とした文俳のことであ た形の文む), 紙文書をコンピュータ上の文替で忠実に の文呰画像認識装置は、文字や画像を統一的に記述する ことができるページ記述 育語を用いて文替を生成するも のである。なお、実施の形態1の文書画像認識装置にお いては、ページ記述首語として、PostScript ent Format:アドビシステム社が公開してい [0013] ここで、流通および閲覧を考慮した文書と できるように、オリジナルの紙文掛が持つレイアウト情 る (紙文售のメタファとしてのWYSIWYGを保存し ptを用いて生成した文樹のことをPostScrlp を用いることにする。以下では、このPostScrl t 文盤と定数することにする。また、PostScri

[00i4]また、利用を考慮した文母とは、オリジナルの紙文盤が持つレイアウトにとらわれず、文替の内容を優先したコンピュータならではの形を持った文替のことである。このように、レイアウトよりも内容を優先した文母を生成するために、実施の形成1の文母画像影響装置は、SGML(Standard Generalized Markup Langage)やHTML(HypertextMarkup Langag

e)、XML (Extensible Markup Langage) 等の構造化配送音節で表現することに よって文璧を生成し、文字や画像が混在した紙文哲をハイバーテキスト化するものである。なお、実施の形態1 の文母画像認識装置においては、構造化配送音節として、HTMLを用いることにする。以下では、このHT MLを用いて生成した文母のことをHTML文章と定義することにする。

Script文書およびHTML文書を生成する文譽画 a, 105b (以下これらを「文書画像処理サーバ10 5」と記述する)で生成した文档を入力して、データベ 一スに格納し、クライアント107からの検索要求に応 じて、抜当する文事を検索して出力する検索サーバ10 8と、から構成されている。なお、109は、LAN等 [0015] 図1は、実施の形態1の文書画像認識装置 のシステム構成を示す構成図である。図1に示す文書画 **像認識装置は、紙文書を光学的に読み取って文母画像を** 生成するカラースキャナ100,モノクロスキャナ10 1およびネットワークスキャナ102と、ファクシミリ 装置103から送信された文書画像を受信するファック キャナ101, ネットワークスキャナ102およびファ ックスモデム104から文蟄頭像を入力し、入力した文 **暋画像に基づいて、PostScript文告およびH** 低文書を光学的に読み取って文書画像を生成するディジ タル複合機106と、ディジタル複合機106から文格 画像を入力し、入力した文書画像に碁づいて、Post スモデム104と、カラースキャナ100,モノクロス TML文書を生成する文書画像処理サーバ105aと、 像処理サーバ105bと、文書画像処理サーバ105 のネットワークを示している。 8 20

[0016] 図2は、図1に示した文語画像認識数置の 概念協成図である。実施の形態1の次軸画像認識装置 は、大きく文格画像入力第200, 前処理時201, 始 組由出部204, 文容生成部210およびデータベース 第213から格成される。なお、文母画像入力部20 0, 前処理部201, 情報由出部204および文母生成 第210は、図1に示した文母画像処理サーバ105に 数当し、データベース部213は、図1に示した検察サーバ108に裁当する。以下に上記各部の結成について 級別する。

【0017】(1) 文塾画像入力部 文書画像入力部200は、カラースキャナ100, モノ

20

るファイル形式)形式を用いることもできる。

クロスキャナ101, ネットワークスキャナ102およびディジタル複合機106で生成した文格画像やファックスモデム104で受信した文华画像を入力するものである。また、文母画像入力部200は、ワードプロセッサ等のアプリケーションプログラムで作成した文母ファイルを入力することもできる。

が処理部201は、文藝画像入力部200を小して文替画像を入力し、入力した文替画像から孤立点ノイズを除去するノイズ除去部202と、入力した文む画像が何い「ているような場合に、傾きを補正するスキュー補正部203とを有している。この前処理部201は、文藝画像についてノイズ除去および傾き補正を行うことにより、後に行われる領域分割処理や文字認識処理において悪影響を及ぼす要因を除去するものである。なお、入力した文藝画像がフルカラーであるような場合、領域分割処理等を容易に、かつ高速に行うことができるようにするため、フルカラーの文藝画像を2値化する2値化処理を行うことにしても良い。

[0019] (3) 情報加出部

た文字領域の文字列について文字認識処理を行うOCR 字領域の文字列のフォントが強闘系 (中ゴシック等) や 非強関系(細明朝等)のいずれであるかを識別する処理 結果に基づいて、文書画像のレイアウトを解析すると共 に、文容の論理的な構造を解析するレイアウト・論理解 び/または図,表,写真等の画像を含む画像領域を認識 ーコラムのいずれで構成されているかを職別する処理お よびセンタリング領域を検知する処理等のレイアウト情 限抽出処理を行う領域分割・僣報抽出部205と、領域 分割・情報抽出部205で分割した画像領域に表が含ま に、枠内の文字領域を抽出する表処理部206と、領域 分割・情報抽出部205および表処理部206で分割し 路207と、領域分割・情報抽出部205で分割した文 を行うフォント韓別部208と、上記各部による処理の カレ、入力した文哲画像から文字列を含む文字領域およ して分割する処理を行うと共に、文魯画像がいかなる段 組循類、例えば、シングルコラム, マルチコラム, フリ **情報抽出部204は、前処理部201から文書画像を入** れている場合に、表の枠と野級の構造を抽出すると共 **枦部209と、を有している。** 

次盤生成路210は、OCR路207による文字認識処理の結果およびレイアウト・論理解析部209による解析が理の結果に基づいて、PostScript文書を生成するPostScript文書を上版するPostScript文音を生成するHTML文書を生成するHTML文書を生成するHTML文書を上による。★1-Tハス

[0020] (4) 文魯生成部

[0021] (5) データペース (DB) データペース213は、PostScript文皆生成 邸211で生成されたPostScript文哲を格納

するPostScript文使DB214と、HTML 文母生成部212で生成されたHTML文部を格制する HTML文母DB215とを有している。 [0022] 次に、上述した構成を有する文書画像認識 装留の動作について、詳細に説明する。図3は、文書画 像認識装置の動作年願を示すフローチャートである。

[0023] 文母回像入力部200は、カラースキャナ100、モノクロスキャナ101、ネットワークスキャナ102名よびディジタル複合数106で生成した文母10層像、並びにファックスモデム104で受信した文母回像を入力する(S301)。また、クライアント107から文母画像やワードプロセッサ等のアブリケーションプログラムで作成した文母ファイルを入力することもで

[0024] 図4は、文登画像入力第200を介して入力した文卷画像を画面表示した様子の一例を示す説明図である。図4において、400は文卷画像認識処理の制御画面を、401は文藝画像入方第200を介して入力した文巻画像を表示する表示画面をそれぞれ示してい

[0025] 図4に示す制御画面400では、以下に説明する各種処理の状行の指定および処理の詳細な設定を行うことができると共に、PostScrlpt次むおよびHTML文むの両方またはいずれか一方の生成を指定することができる。なお、図4においては、文母回位を入力し、各種処理の実行を指定することができる制御画面400を示したが、予め設定した条件に基づいて、文母の生成・格納まで自動的に実行できるようにすることもできる。

[0026] 続いて、ノイズ除去部202は、文砂画像 入力部200を介して文好画像を入力し、入力した文財 画像から孤立点ノイズを除去する(S302)。また、 入力した文哲画像が飼いているような場合、スキュー制 正部203は、文母画像の何きを描正する(S30 10027] 関域分割・債報抽出第205は、ノイズ除去第202およびスキュー和正部203からなる前処理第201から文母回像を入力し、入力した文母回像から文字列を含む文字のはおよび/または図、我、写真等のの画像を含む回像前域を認識して分割する処理を行う(S304)。分割した各領域には、文字領域か、回像領域の場合にはさらに図、我、写真が等の領域が、回像領域の場合にはさらに図、我、写真が等の領域の組額および領域の位置が同程情報として付与される。[0028] 領域分割処理を行った後、領域分割・指報抽出処理第205は、文書画像がいかなる段組電額、例えば、シングルコラム、マルチコラム、フリーコラムのいずれで構成されているかを整別する処理およびセンタリング領域の核知処理等のレイアウト情報を抽出する処

理を行う (S305)。 [0029] 具体的には、文字節域同士の間の距離(空

22

[0030] その後、表処理部206は、領域分割・情 6)。この判定は、上記のように装置側で自動的に行う 白部分)および野椒を検出し、検出した距離および野椒 の本数と共に、領域分割処理で分割した文字領域の位置 て、表を含む領域が存在するか否かを判定する(S30 ことにしても良いし、ユーザが指定しても良い。 表が含 に関する原性情報に基づいて、段組種類の判定を行う。 報抽出部205で分割した各質域の属性情報に基づい まれていない場合には、ステップS308に進む。

02 文字領域を加出する(S307),このように、表の中 [0032] そして、OCR部207は、領域分割・情 報加出部205および表処理部206で分割した文字領 すなわち、OCR部207は、文字領域について、行切 協々の文字パターンについて文字認識処理を行う。加え [0031] 一方、表が含まれている場合、表処理部2 06は、表の枠と罫線の構造を抽出すると共に、枠内の から文字質域を抽出することにより、次のOCR部20 り出しおよび文字切り出し処理を行い、文字切り出した て、〇CR部207は、文字認識結果に対して、言語処 域の文字列について文字認識処理を行う(5308)。 7 において、表中の文字認識処理を行うことができる。 型による観り補正を行う。

[0033]フォント韓別部208は、領域分割・情報 抽出部205で分割した文字領域の文字列について、行 (S309)。 具体的には、例えば、照画幕密度やラン レングスの分布等に基づいて、フォントの特徴を讃別す [0034] 続いて、レイアウト・வ理解析部209 単位でフォントが強闘系(中ゴシック等)や非強闘系 (御明朝等) のいずれであるかを散別する処理を行う

30 は、上記各部による処理の結果に基づいて、文整画像の レイアウトを解析する (S310). ここで行われるレ イアウトの解析処型には、例えば、タイトル部,小見出 し邸、キャプション、ヘッダ・フッタ部の検出処理が含

\$ オント説別部208による設別結果を用いて、タイトル とはサイズや行ピッチが異なり、また、存在する位置も **水文とは若干離れていることから、領域分割・情報抽出** 部205で付与した領域の位置に関する属性情報や、フ [0035] ここで、タイトルは、一般的に本文の文字

部を校出することができる。

トが、同一の文字領域中の他の文字のものと異なる場合 [0036] 小見出しは、本文の文字と文字サイズがほ ぼ等しい場合も多く、本文に近接した場所に位置するこ そこで、各文字領域の先頭行の文字サイズまたはフォン とから、本文と同一の領域に存在していることも多い。 に、先頭行を小見出し行と判定する。

S [0037]また、キャプションは、図, 表, 写真等の 回像に付与されたものであり、一般的に回像領域の近傍 で、本文とは離れた位置に存在すること、さらには、

「図8」等の図や表等を指し示す師が含まれていること から、これらの条件を満たす行をキャブションと判定す る。そして、キャプションと対応する図との間にリンク [0038] さらに、ヘッダ・フッタは、文書画像の上 娩出する。具体的には、例えば、領域散別・情報抽出部 でシングルコラムと識別された場合、センタリング行よ ヘッダの行のレベルについても磔別する。また、マルチ コラムと蟄別された場合、段に属しない上下の行がヘッ **ドに存在するため、該当する行をヘッダ・フッタとして** り上の行がヘッダ部と判定することができる。この際、 グ・フッタ部となる。

は、文協の論理的な構造を解析する (5311)。 具体 [0039] その後、レイアウト・簡単解析部209

的には、図、表、写真等に付与されたキャプション中の 「図8」等の語と同一の語を本文中から見つけ出し、こ れらの文字間にリンクを生成する処理を行う。 この処理 においては、〇CR部207による文字認識の結果を用 いるため、各文字パターンの候補文字の全てを使用する [0040] そして、文徴生成部210は、上述した情 文書を生成する処理を行う (S312)。 すなわち、P ことによって、OCRエラーに対処することができる。 報抽出部204の各部で行った処理の結果に基づいて、

字認識確信度の低い文字については、文字認識の結果で 04で行った文字認識やレイアウト解析等の結果に基づ いて、PostScriptで各ページを表現した文哲 を生成する。なお、OCRエラーをカバーするため、文 ostScript文盘生成邸211は、情報抽出部2 はなく、元の画像を用いることができる。

[0041]また、HTML文掛生成部212は、情報 **抽出部204で行った文字認識やレイアウト解析等の結** 果に基づいて、タイトル、殷務等を示すタグを各文字列 に付与し、HTML文母を生成する。

[0042]その後、PostScript文書はPo 3)。なお、同一の文替画像から生成したPostSc stScript文俳DB214に、HTML文俳はH は、一方から他方を呼び出すことができるように、相互 IML文書DB215にそれぞれ格納される(S31 r ipt文書とHTML文書の両方を登録する場合に に関連づけを行うことにしても良い。

示することができる。すなわち、クライアント107か [0043] なお、PostScript文費DB21 4 およびH T M L 文書 D B 2 1 5 にそれぞれ格納された PostScript文書およびHTML文書は、検案 処理を行うことによって各クライアント107で回面投 らの複繁要状に応じて、複鑑サーバ108がPostS cript文費DB214およびHTML文費DB21 07は、検索サーバ108から検索結果を入力し、検索 5から該当する文書を検索して出力し、クライアント1 要求に該当する文書を画面表示する。

表示した様子は元の紙文告とほぼ同一であるが、内部情 街と同一のレイアウトの文音を容易に生成することがで きる。すなわち、紙文掛を再現した文費を生成すること とができるような、黄通および閲覧に適した文뿁を得る ことができる。なお、PostScript文費を画面 報は保持されているため、検索を行ったり、再利用した [0044]図5は、PostScript文書を画面 技示した様子の一例を示す説明図である。図5に示すよ l p t で表現した文沓を生成することにより、元の紙文 うに、図4に示した文眷画像に基づいてPostScr により、文書作成者の意図をできるだけ正確に伝えるこ りすることができる。

た様子の一例を示す説明図である。図6は、図と図番号 のハイパーテキスト化を行ったものであり、例えば、本 に該当する図が画面表示される。このように、紙文費が 持つレイアウトにとらわれることなく、ハイパーテキス ト化することにより、紙文甞の内容を優先したコンピュ [0045]また、図6は、HTML文哲を画面投示し 文中の「図9」をマウス等でクリックすると、「図9」 ータならではの形を持った文書を生成することができ

装置によれば、文街画像から文字コードを抽出するとい 持つ様々な情報を抽出して利用するため、紙文聲の再現 [0046] このように、実施の形態1の文哲画像認識 う単なる文字認識処理を行うだけではなく、文啓画像の を優先した文書や紙文書の内容を重視した文書等、利用 目的に広じた様々な形態のコンピュータ上で利用可能な 文格を生成することができる。

9を介したシステムとして実施の形態1の文街画像認識 [0047] なお、図1においては、ネットワーク10 装置の構成を説明したが、図2に示す機能を1台のコン ピュータに持たせることにより、スタンドアローンの形 **態で文誓画像認識装置を構成することもできる。** 

【0048】また、実施の形態1の文譽画像認識装置で は、上述したPostScript文書やHTML文書 を生成することにしたが、これらに限定するものではな く、必要に応じて他の形式の文徴を生成することもでき

認数装置は、実施の形態1で説明したようにして生成し たPostScript文俳やHTML文俳を効率良く 検索することができるようにしたものである。 具体的に は、OCR部207で文字認識した文字列から所定の文 字列を抽出し、抽出した文字列を対応するPostSc rlpt文書やHTML文書に関連づけておき、該当す る文字列を検索することにより、関連づけされたPos tScript文盤やHTML文替を検索結果として出 力できるようにするものである。以下では、この文字列 【0049】 (実施の形態2) 実施の形態2の文铅画像 のことをキーテキストと定義することにする。

ಜ [0050] 図7は、実施の形態2の文语画像認識装置

t文音やHTML文告を検索結果として出力する。

特開平11-66196

9

の概念構成図である。図7において、実施の形態1で説

明した図2と同一の構成については同一の符号を付すこ とにし、それらの辞細な説明については省略する。

Cript文替またはHTML文格を検索結果として出 に示すように、OCR部207で文字認識した文字列か **ーテキスト登録船101と、検索要求を入力し、キーテ** て、核当するキーテキストに関連づけられたPostS ら上近したキーテキストを抽出するキーテキスト抽出部 700と、キーテキスト抽出部700で抽出したキーテ [0051] 実施の形態2の文徴画像認識装置は、図7 キストを入力し、キーテキストDB102に登録するキ キストDB702に登録されたキーテキストを検索し カする検索処理部703と、を備えている。

02は、検索処理部703に設けられる。この検索処理 索処理を行う。また、実施の形態2の文掛画像認識装置 クライアント107からの検索要求に基づいて、上記検 をスタンドアローンの形態で構成した場合には、直接検 [0052] なお、PostScript文書DB21 4, HTML文告DB215およびキーテキストDB7 部703は、図1における検索サーバ108に該当し、 森要求を入力して検索処理を行う。

20

[0053] 上記キーテキスト相出部207で抽出する る文字列や、図番を含むセンテンス, このセンテンスを トとして抽出しても良い。なお、上配キーテキストを抽 出するには、文锴画像のレイアウトを解析する必要があ キーテキストとしては、文母を端的に扱した文字列、例 えば、文書全体、章、節のタイトルや、ヘッダ・フッタ 文掛中の図等を基準として、図のキャブションを構成す 含むパラグラフおよびページ単位の文字列をキーテキス ることから、キーテキスト協出館700は、レイアウト ・飴理解析部209による解析結果を用いて、キーテキ 等の否誌的事項、文俳の要約文等が考えられる。また、 ストの抽出処理を行うようにしても良い。

[0054]また、キーテキスト抽出部701は、上記 力し、入力したキーテキストを文铅生成部210で生成 したPostScript文事やHTML文母に関連づ キーテキスト抽出部700で抽出したキーテキストを入 け、キーテキストDB702に格納する。

トDB702から該当するキーテキストを検索する検索 04は、検索要求を入力して検索エンジン705に検索 を入出力部704に出力する。入出力部704は、入力 【0055】さらに、検索処理部103は、検案要求を と、入出力部704から検索要求を入力し、キーテキス 要求を出力する。検察エンジン705は、入出力部70 4から検索要求を入力し、キーテキストDB702から 核当するキーテキストを検索し、核当するキーテキスト したキーテキストに関連づけられたPostScrlp エンジン705とを有している。具体的に、入出力部7 入力すると共に、検索結果を出力する入出力部704 49

特開平11-66196

6

(0056)このように、実施の形態2の文母回像認識 数個によれば、文書画像中の文字列をキーテキストとし て文むの検索を行うことにしたため、検索要求に対し て、取も適切な検索結果を持ることができると共に、高 [0057]なお、上述した英施の形態2においては、キーテキストを用いて文替の校案を用いることにしたが、キーテキストに代え、OCR的207で文字認識した結果全てを用いて全文検察を行うようにすることもできる。図8は、実施の形態2の文哲画検認識装置の変形

例を示す概念構成図である。

**遠な検索処理を実現することができる。** 

(0058) 図8に示すように、上述したキーテキスト 市出部700, キーテキスト登録的701およびキーテ キストDB702に代えて、テキスト登録的800およ び全文検索用テキストDB801を設け、テキスト登録 断800がOCR第207から文字認識結果、即ち文苷 国像中のテキストの全文を入力し、入力したテキスト 文仿生成第210で生成したPostScript文 をHTML文音を同述づけし、全文検索用テキストDB 801に登録されたテキストを用いて、PostScript文部 1pt文音やHTML文音を検索用テキストDB 801に登録されたテキストを用いて、PostScript文字を の場合は、テキストを全文検索用テキストDB 801に登録されたテキストを用いて、PostScr ウ場合は、テキストを全文検索用テキストDB 01に登録するため、検索の仮に各ファイルのオープン・クロ ーズという処理が不受となるため、検索処理の高速化を 図ることができる。

2

[0059] (支施の形態3) 続いて、実施の形態3の文体回像20数数とついて説明する。実施の形態3の文体回像20微数阻は、実施の形態1のものと同様に、減過および間質を考慮した文曲と、利用を考慮した文章を生成することができるようにしたものである。

[0060] 契縮の形態3の文母回像認識装置における 流過および閲覧を考慮した文母とは、オリジナルの紙文 母を読み取って生成した文帝回像であり、ここではイメ ージデータと定義することにする。また、利用を考慮し た文母とは、文母回像中の文字列について文字認識を行 った結果であり、ここではテキストデータと定義するこ とにする。

[0061]図9は、実施の形倣3の文華画像認識装置の概念構成図である。なお、図9において、実施の形態 401で図明した特成と同一の結成については同一の符号を付し、これらの詳細な説明は省略する。

【0062】実施の形態3の次母面像認識装置は、大きく文母面像入力部200、前処理部201、情報抽出部204、登録部900およびデータベース部213から構成される。なお、文母面像入力部200、前処理部201、情報抽出部204および登録路900は、図1に示した文母面像処理サーバ105に該当し、データベース部903は、図1に示した必要が一バ108に該当する第903は、図1に元した必要が

[0063] 文哲画像入力節2004よび前処理第201については炎髄の形態1で設明した適りである。 また、情報加出節204においては、PostScrlpt文章やHTML文章を生成しないため、フォント観別的2084よびレイアウト・福理解が第209が省略されている以外は実筋の形態1で設明した適りである。

[0064]実施の形態1の次母回像認識装置と異なるのは、イメージDB登録901およびデキストDB登段的901およびデキストDB登段的902、イメージDB904およびテキストDB905を悩えたデータベース部94およびテキストDB905を悩えたデータベース部9

[0065]イメージDB登録部901は、前処理部201からノイズ路去およびスキュー植正を行った後のイメージデータを入力し、入力したイメージデータをデータベース部903のイメージレB904に登録する処理を行い、テキストDB登録部902は、〇CR路207文文認識処理を行った結果であるテキストデータをデータベース部903のテキストDB905に登録する処理を行う。なお、同一文曲のイメージデータとテキストデータの両方を登録する場合には、一方から他方を呼び出すことができるように、相互に関連づけを行うことにしても良い。

[0066] なお、イメージDB904およびテキストDB905にそれぞれ格納されたイメージデータおよびテキストデータは、検索処理を行うことによって各クライアント107で回面表示することができる。すなわち、クライアント107からの検察要求に応じて、検索サーバ108がイメージDB904およびテキストDB905から核楽して出力し、クライアン80ト107は、検索サーバ108から検索結果を入力し、検索要求に該当する文語を適面表示する。

 [0068] (契縮の形盤4) 実施の形態4の交替回像 認識装置は、実施の形態3で説明したようにして生成したイメージデータやテキストデータを、実施の形態2で改明したキーデキストを用いて効率良く検索することができるようにしたものである。

[0069]図10は、実施の形態4の文書画像影響装置の概念格成図である。図10において、実施の形態1で設明した図2および実施の形態2で説明した図7と同一の格成については同一の符号を付すことにし、それらの辞題な説明については名略する。

【0070】実施の形態もの文母画像認識装置は、図10に示すように、文母画像のレイアウトを解析し、OCR8207でで実践難した文字列から上述したキーテキ

ストを抽出するキーテキスト抽出筋700と、キーテキスト抽出筋700で抽出のスキーテキストを入力し、キーテキストDB702に登録するキーテキスト登録的701と、直接に検索要求を入力し、キーテキスト登録の702に登録されたキーテキストを検察して、該当するキーテキストに関連づけられたイメージデータやテキストデータを検索結果として出力する検索処理第703と、を備えている。

(0071)上記キーテキスト抽出部207で抽出するキーテキストとしては、文章を始めに表した文字列、例えば、文事全体、章、節のタイトルや、ヘッグ・フッグ等の心は的事項、文章の要約文等が考えられる。また、文章中の図等を基準として、図のキャブションを構成する文字列や、図番を含むセンテンス、このセンデンスを含むパラグラフおよびページ単位の文字列をキーテキストとして抽出しても良い。

[0072]また、キーテキスト抽出部701は、上記キーテキスト抽出部700で抽出したキーテキストを入力し、対応するイメージデータやテキストデータに関連づけ、キーテキストDB702に格納する。

【0073】さらに、検索処理部703は、検索要求を 入力すると共に、検索結果を出力する入出力部704 と、入出力部704から検索要求を入力し、キーデキストDB702から該当するキーデキストを検索する検索 エンジン705とを有している。具体的に、入出力部704は、検索要求を入力して検索エンジン705に検察 94は、検索要求を入力して検索エンジン705は、入出力部704から検索エンジン705は、入出力部704に対する。検索エンジン705は、入出力部704に対策のキーデキストDB702から検当するキーデキストを入出力第704に対方。入出力第704は、入力したキーデキストに関連づけられたイメージデータやデキストデータを検索結果として出力する。

[0800]

[0074]上記檢案処理部703は、図1における検索サーバ108に該当し、クライアント107からの検索要求に基づいて、検索処理を行う。また、装飾の形態2の文書画像認識装置をスタンドアローンの形態で特点した場合には、直接検索要求を入力して検索処理を行

【0075】このように、実施の形態4の文書画像影談数盤によれば、文書画像中の文字列をキーテキストとして文書の検索を行うことにしたため、検索要求に対して、最も適切な検索結果を得ることができると共に、高速な検索処理を実現することができる。

[0076]なお、上述した実施の形盤もにおいては、キーテキストを用いて文母の検索を用いることにしたが、キーテキストに代え、OCR邸207で文字認識した結果全てを用いて全文検索を行うようにすることもできる。図11は、実施の形態もの文母画像認識技器の変形例を示す概念構成図である。

[0077]図11に示すように、テキスト登録部80 50

0 および全文検索用テキストDB801を設け、テキスト登録器 8 0 0 がOC R 間 2 0 7 から文字認識結果、即ち文書の場合のテキストの全文を入力し、対応するイメージデータやテキストアークに関連づけし、全文検索用テキストDB801に登録する。その結果、全文検索用テキストDB801に登録されたテキストを旧いて、イメージデータやテキストデータを検索することができる。この場合は、テキストを全文検索用テキストDB801に登録するため、検索の度に各ファイルのオープ・クローズという処理が不迟となるため、検索の理の高

強化を図ることができる。 【0078】なお、上述した攻筋の形態1~4の文语面像段識装置は、それぞれ任意に組み合わせて1つの文语画像段識装置を構成することができる。

[0079] さらに、上述した実施の形態1~4の文母 画像認識装置としてコンピュータを機能させるプログラムを作成し、これらをハードディスク, フロッピーディスク, CD-ROM, MO, DVD等のコンピュータ説 み取り可能な記録媒体に記録して、記録媒体を介してプログラムを配布することができる。そして、記録媒体に記録されたプログラムをコンピュータで読み出して実行することにより、上述した文格画像認識装置を実現することができる。

ន

字認識手段による文字認識結果およびレイアウト情報抽 化記述 育語を用いた第2の文督を生成する第2の文事生 単なる文字認識処理を行うだけではなく、文啓画像の持 つば々な情報を抽出して利用することにより、紙文母の [発明の効果] 以上説明したように、本発明の文哲画像 取ることによって生成した文書画像を入力する入力手段 と、入力手段を介して入力した文街画像から文字列を含 む文字領域および/または図、表、写真等の画像を含む 画像領域を認識して抽出する領域抽出手段と、領域認識 手段で抽出した文字領域の文字列について文字認識処理 を行う文字認識手段と、領域加出手段の抽出結果に払づ いて、文書画像のレイアウトを解析し、レイアウト情報 を抽出するレイアウト情報抽出手段と、文字認識手段に よる文字認識結果およびレイアウト情報抽出手段による レイアウト情報抽出結果に基づいて、ページ記述質節を 用いた第1の文書を生成する第1の文書生成手段と、文 出手段によるレイアウト情報抽出結果に基づいて、構造 成手段と、第1および第2の文掛生成手段で生成した第 1 および第2の文書をそれぞれ格納する格納手段と、を 崩えたため、文格画像から文字コードを抽出するという 利用目的に応じた様々な形態のコンピュータ上で利用可 忍離装置 (請求項1) によれば、紙文盤を光学的に読み **再現を優先した文むや紙文むの内容を取視した文俳等、** \$ 8

能な文書を生成することができる。 [0081]また、本発明の文母画像認識装置(結氷項 2)によれば、請求項1に記載の文母画像認識装置にお 9

[図]

図である. 103 101 2 コンピュータならではの形を持った文替を生成すること PDF形式によって表現された文替であるため、文쑙作 3) によれば、請求項1に記載の文番画像認識装置にお いて、第2の文むは、SGML, HTMLまたはXML アウトにとらわれることなく、紅文俳の内容を優先した いて、第1の文書は、PostScript形式または 成者の意図をできるだけ正確に伝えることができるよう [0082] また、本発明の文書画像認識装置(間求項 によって表現された文晳であるため、紅文街が持つレイ な、试過および閲覧に適した文樹を得ることができる。

コンピュータ上で利用可能な文봡を生成することができ 20 内容を重視した文書等、利用目的に応じた様々な形態の [0083] さらに、本発明のコンピュータ説み取り可 他な記録媒体(請求項4)によれば、請求項1~3のい ずれか1つに記憶の文替画像認識装置の各手段としてコ ンピュータを徴能させるためのプログラムを記録したた め、記録したプログラムをコンピュータ上で実行するこ とにより、文藝画像の持つ様々な僧報を抽出して利用す ることにより、抵文告の再現を優先した文書や低文帯の

【図1】 実施の形態1の文書画像認識装置のシステム格 [図面の簡単な説明]

【図2】図1に示す文書画像認識装置の概念構成図であ 成を示す構成図である。

【図4】 実施の形態1の文替画像認識装置において、文 【図3】 実施の形態1の文書画像認識装置の動作手順を 示すフローチャートである。

**铅画像入力部を介して入力した文铅画像を画面表示した** 【図5】 実施の形態1の文書画像認識装置において、生 様子の一例を示す説明図である。

2

成したPostScript文掛を画面投示した様子の 【図6】 東筋の形態1の文魯画像認識装置において、生 一例を示す説明図である。

成したHTML文書を画面表示した様子の一例を示す説 [図7] 玫鉱の形態2の画像認識装置の概念構成図であ

40 【図8】 収施の形態2の文書画像認識装置の変形例を示 す既急指成図である。

[図9] 実施の形態3の文書画像認識装置の概念構成図

【図10】 実施の形態4の文魯画像認識装置の概念構成

[図11] 実施の形態4の文魯画像認識装置の変形例を

ネットワークスキャナ モノクロスキャナ ファクシミリ装置 カラースキャナ ドす概念構成図である。 [符号の説明] 102

文告回像処理サー ファックスモデム

105 (105a, 105b)

ディジタル複合機 **食素サーバ108** クライアント ネットワーク 901 0 8 601 0 7

2番画像入力部 加加西部 200 201

スキュー植正部 イズ除去部 202 203

領域分割・情報抽出部 資報抽出部 205 204

麦処理部 OCR思 206 207

フォント類別部 208

PostScriptX俄DB HTML文魯生成部 データペース部 213

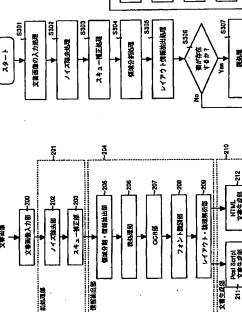
キーテキスト抽出部 キーテキスト登録部 HTML文魯DB 215 00 0.

--テキストDB 设索処理部 八出力部 0 2 03 0.4

全文検索用テキストDB テキスト登録部 **検索 エンジン** 800 801 7 0 5

ŝ

[図3] [図2]



レキント観別処理

フィアシャ無控対路 文字四篇处理 微知道

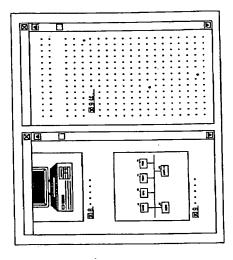
文書生成処理

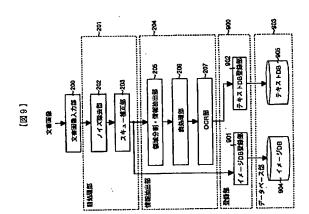
抽場解析地理

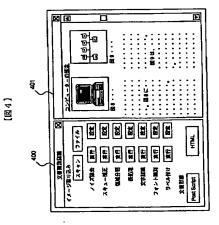
文書格數処理

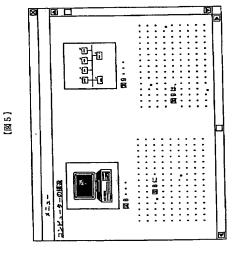
H Y

[9図]

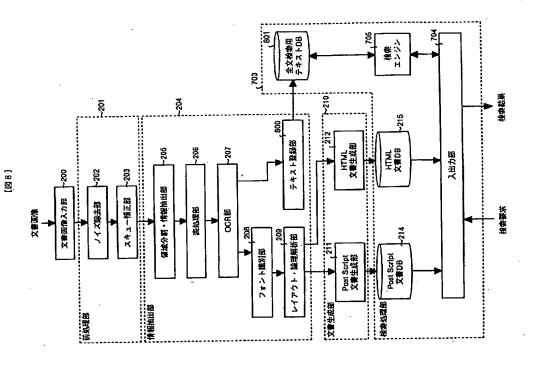


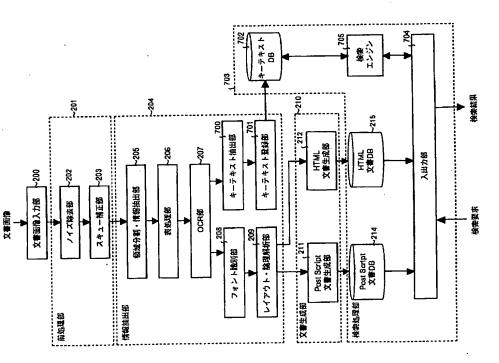




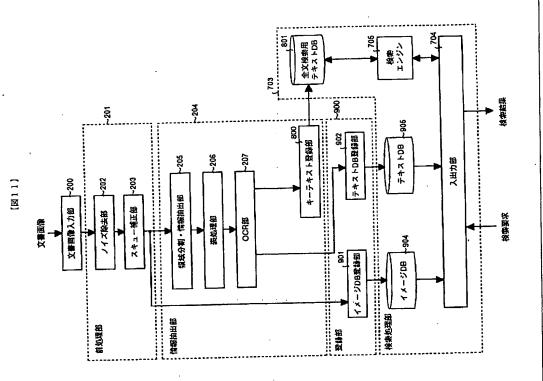


[図7]





[図10]



F1 H04N 1/40

数四配布

(51) Int. Cl. a H 0 4 N 1/40

フロントページの統合

